DOCUMENT RESUME

ED 367 156 FL 021 878

AUTHOR Weir, C. J.; Roberts, J.
TITLE Evaluating a Teacher Training Project in Difficult Circumstances.
PUB DATE 91
NOTE 20p.; In: Anivan, Sarinee, Ed. Issues in Language Programme Evaluation in the 1990's. Anthology Series 27; see FL 021 869.
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English (Second Language); Foreign Countries; Grade 8; Language Teachers; Program Effectiveness; *Program Evaluation; *Research Methodology; Sampling; Secondary Education; Second Language Instruction; *Second Language Programs; *Teacher Education Curriculum; *Testing Problems; Work Environment
IDENTIFIERS *Nepal

ABSTRACT

A recent evaluation of the effect of a Nepalese language teacher training program on student language learning is described and discussed, focusing on the difficulties faced by external evaluators working in difficult circumstances and not on study results. The program provided inservice training for secondary school teachers of English as a Second Language, using locally trained Nepali staff and an expatriate training officer. The evaluation study involved comparison of language performance of students taught by program-trained teachers with that of students taught by untrained teachers. Baseline measurements of student performance were taken at grade 8 in 12 experimental and 12 control schools, and posttest results were ultimately available from 11 experimental and 11 control schools. In addition, teachers were observed in class, were interviewed, wrote reports of their lessons, and submitted samples of pupils' work. Selection of participating schools was heavily influenced by serious problems of communication and information gathering in Nepal. Methodological problems were found in the following areas: the design of language tests used; quantity and reliability of data gathered; potential for bias in the observation structure; use of external evaluators; and sampling. Some recommendations are made for improving the process. (MSE)

EVALUATING A TEACHER TRAINING PROJECT IN DIFFICULT CIRCUMSTANCES

C J Weir and J Roberts

This paper presents a description of the procedures adopted in a recent evaluation of the effect of a teacher training programme on student language performance. It is hoped that the account will lead to constructive discussion of how to improve the methodology employed. It considers the problems that may be faced by external evaluators working in difficult circumstances. We are grateful to the Overseas Development Administration in the United Kingdom for giving us permission to report our methodology.

1 BACKGROUND

The SEPELT Inset Project in Nepal was set up to provide 1080 standard 8-10 (Upper Secondary Level) English teachers with one month's inservice training, delivered by locally trained Nepali staff, working from a standard course manual and supported by an expatriate training officer. It ran from 1987 to 1989.

The long term goal of the training was to improve students' performance in the School Leaving Certificate English examination. The course provided training in basic ELT procedures designed to enhance the teaching of the National English Curriculum.

The Nepal baseline study described in this paper was a small scale, field based, non equivalent group study, contrasting the learning gains of students in the grade 8 classes of 11 trained teachers and 11 untrained teachers. The study established procedures for measuring the effect of the SEPELT training on students' language performance. It was also concerned with determining the suitability of these procedures for evaluating similar projects elsewhere.

A small scale non-equivalent control group pretest- posttest design was employed. In this design two groups of students which are similar, but which are not formed by random assignment, are measured both before and after one of the groups undergoes the experimental treatment.

In this case the experimental treatment took the form of instruction by teachers who had attended the SEPELT training course. We were concerned to see if, with faithful implementation of the training, there would be superior learning gains by this group as evidenced by improvement in student language test scores.

As well as testing students we had to monitor the performance of trained (Experimental) and untrained (Control) teachers to establish that the treatments received by the pupils were indeed different, i.e., were our control and experimental groups exposed to different language instruction?

We made short visits to Nepal in November 1988, January 1989 and November 1989. As a result of the first visit the baseline framework was set up and technical staff contracted for data collection (The New Era Research organisation). A short training course was provided for technical staff during the second visit. The final visit was made in order to monitor data collection.

2 The Language Test Instruments

To determine the effects of the training programme, base line tests were administered to the new intakes in grade 8 at a selected group of schools (12 Control and 12 Experimental in the first instance). This took place at the start of the school year in February 1989. These classes were of both the experimental type, where the teachers had been on a training course

and of the control type, where the teachers had received no prior EFL training. The tests were readministered at the end of grade 8 in November 1989.

Part I of the battery was constructed to sample as widely as possible the structural elements in the English syllabuses for years 7 and 8 in the Nepali Upper Secondary School System on the basis that such linguistic elements would be accessible to both control and experimental groups. The two parts of this general proficiency section of the battery were prepared in advance of the November 1988 visit and piloted during that visit.

They consisted of:

PART IA Selective Deletion Gap Filling 100 items 1 hour

Passsages were selected and reconstructed from the grade 7 and 8 textbooks used in Nepal. They were rewritten taking care to employ only those structures and lexical items occurring in these course books. Items were then deleted from these passages to sample as far as possible the range of structural items in the national curriculum for these grades. The task for students was to repair the deleted items by writing the missing word on an Answer Sheet provided. A similar technique was employed in Davies et al's 1984 Survey of English Language Teaching in Nepal. The latter differed in that it was a test which had originally been designed for use in Malaysia.

The properties of many realisations of this test method are high correlations with other general proficiency measures and with tests of reading comprehension in particular. We wanted to have a large number of items with a wide range of difficulty because the test might have to show improvement over a period from 1 to 3 years. The problem with the more familiar passage plus comprehension question format is that the number of items that can be set is restricted. By using a gap filling test it is possible to create a far greater number of items which we thought would demonstrate development in linguistic competence.

It was hoped that students would be able to complete some of the items learnt in Grade 7 in the first administration and by the end of the year it was hoped they would be able to score on those set on the year 8 syllabus.

The experimental group might well be expected to outperform the control group on this section of the test as the training course was aimed at improving the teaching of the existing materials in the Grade 8 reader and in the long term raising the number of successful passes in the SLC in Grade 10. The ability of students to produce structural items was an anticipated result of the trainees' implementation of their training. It would still however be a fair test for the control group as it did not contain any structures or lexis that were not present in books 7 and 8. Gap filling exercises are also present in the course books.

PART IB Dictation 30 minutes.

This was a forty item test, again of general proficiency, based on sentences (mainly imperatives, instructions and directions) taken from books 7 and 8.

It differs from test 1A in that as well as correlating well with other measures of general proficiency it has a good record of correlating highly with other measures of listening ability.

It was felt that if this test was to be readministered in 1989 and 1990 then it obviously must not be too easy to start with or otherwise a ceiling effect might negate the possibility of measuring achievement and identifying any increase in scores in either of the two groups over the period of the study.

As with the gap filling test the purpose was to determine whether the experimental group performance would outdistance the control group particularly as the trained teachers were trained to use more English as against Nepali in the classroom and a number of the activities in the training manual encouraged this. So once again this measure of general proficiency was designed to reflect the main purpose of the training course namely to enable the teachers to improve the effectiveness of their teaching.

As tests of general language proficiency, both dictation and gap filling can still be considered fair to the control group as they do not contain any language extraneous to the course materials and to a lesser or greater extent dictation and gap filling will occur in the lessons of both groups. Dictation is advocated in the national curriculum which states that at the end of Lower Secondary Level the student will be able to "take a dictation from any of the prescribed materials in the text book".

To summarise, the expectation was that with improved teaching methods and a greater use of English in the classroom the experimental group would improve at a greater rate and that eventually performance in the SLC would reflect this.

PART II

Another purpose of the discussions with the trainers during the first visit was to identify criterial, behavioural differences that might be expected to emerge in students' performances as a result of this short in service training programme.

We have already commented on the general aim of the training to make teachers more efficient at what they did already and our feeling was that Part I of the test adequately catered for this aspect. Our concern in Part II was to reflect any differences in kind, in terms of student performances that might be expected to emerge from the training.

The training team's view was that it was in the skill of writing that clear differences and new behaviours were likely to occur. The experimental students were more likely to be able to create meaningful new sentences and to execute controlled writing tasks. The control group were more likely to copy from the board and memorise and reproduce paradigms provided by the teacher. They also felt that there might be increased oral interaction among students as a result of the training. However, the considerable difficulties and vast expense of conducting spoken language exams precluded their use under the conditions obtaining in Nepal.

We restricted ourselves to trying to establish whether any differences in written production occured through employing test tasks to relect these activities in Part II of the battery.

During our first visit in November 1988 we had asked trainers and trainees to write a short essay on a topic that would be familiar and accessible to their students in year 8 and also to prepare a framework for a cued writing task on the subject. We thus had thirty five examples to provide us with an idea of levels, topic areas and content. This enabled us to produce four cued writing tasks from which we selected the final two used in the battery after trialling in Katmandhu in December 1988.

3 The Process Instruments

Development of Instruments

Prior to the November 1988 Nepal visit, an inventory of training characteristics was produced from the manual and other available documentation.

In discussion with Nepali trainers during the November 1988 visit, the features that they considered both to be of highest priority and the best discriminators of trained and untrained teachers were selected from the inventory and the resulting short list was then used as the basis for the observation instrument. Additionally this list helped us identify more clearly those training characteristics potentially capable of effects on measurable pupil performance which would need to be reflected in the tests.

A revised teacher self report instrument was produced after discussion with Nepali trainers, and a copy left with a request for piloting.

On return, we trialled a draft observation schedule in a local secondary school's classes of French. The design of the resulting schedule took into account its implementation by technical staff, with an emphasis on simplicity of use and a focus on low inference observations. The revised form was the basis for observer training conducted in January 1989 by Weir (see Appendix 1 for details of the schedule).

Instruments: Observation

On page one of the schedule details of the school, observer, teacher, class and lesson are recorded. On page two the observer must code three five minute samples of classroom talk into categories of: teacher and pupil talk; use of English and Nepali; use of questions as contrasted with all other forms. The samples are selected according to strict time criteria. These data give an indication of the proportion of English used in class, who uses it and to what extent questions are used by either teachers or pupils. On page three, the observer is provided with a checklist of classroom activities, to indicate if an activity has occurred, irrespective of duration. For each activity identified, a short illustrative note is required. These data should indicate the occurrence of activity types which discriminate trained and untrained teachers. Also, the observer writes unstructured notes to describe the whole lesson in terms of teacher and student activities, including those in progress during coding. While these notes are unstructured, they are based on a previously identified lexicon of appropriate action verbs. These notes provide a "thumbnail sketch" of the observed lessons which can also be cross-referred with checklist and coding data.

On page four, the observer is required to estimate overall talking time in the lesson and the respective proportions of English and Nepali used by teachers and pupils.

It may seem that the data obtained present a somewhat narrow and simplistic account of classroom processes. It should be noted that there is nothing to be gained by making observers' tasks more complex than necessary; that key training and discriminating characteristics can be identified; and that talk in elementary ELT classrooms is intrinsically controlled and restricted in range.

Additional Data Collection

Two other forms of convergent data were required: teachers' self report lesson descriptions and pupils' work.

Self Report

On each visit by technical staff, the teachers were given self report forms. They were asked to describe three recent, typical lessons they had given.

Self report is often considered unreliable, reflecting impression management rather than actual practice. However,teachers are unlikely to report doing what they never do, or what is unknown to them. These data were used, with caution, as additional information on the customary practice of teachers in the two groups.

Pupils' Work

As part of the final observation visit, technical staff were asked to obtain samples of work from about 5 pupils in each class. These data were obtained to help identify discrepancies with observational and self report data.

Teacher Interviews

Structured interviews were held with the 18 teachers who attended a meeting during the third visit. Data were obtained on the following features: years of service in the school; educational background and training; the teacher's place of origin; other occupations; number of pupils in the school; number of pupils in class ; number of lessons per week; number of lessons in the year; school's SLC pass results for last year, both general and in English only; likelihood of teacher continuing with the class in grade 9; an estimate of the teachers' level of oral English, using the British Council ELTS scale.

4 Selection of sample

Location

The design of the study was heavily influenced by the serious problems of communication and information gathering in Nepal. Most schools do not have telephones and postal delivery is highly unreliable. Telegrams can take up to 6 weeks to arrive and the only means of ensuring messages getting through is by personal delivery. These problems were compounded in November 1989 by India's obstruction of key Nepali imports, notably petrol, which made all travel extremely difficult. District Education Office files often do not contain complete or up to date information, such as lists of school staff. Without actually visiting the schools it is not possible to ascertain whether particular teachers are still teaching there or not. As a result a large scale study of a widely dispersed sample of teachers was always out of the question. These factors also indicated the need to employ experienced field workers through the local New Era research organisation rather than ELT subject specialists with no fieldwork experience.

Kathmandhu valley would have been by far the most convenient place to conduct the study, but it is evident that it is quite unlike any other region of Nepal, and would not have provided representative sample schools, particularly since most of the training took place outside Katmandhu valley.

Pokhara region was chosen after discussions with the project leader during the summer of 1988. It was considered to be a fair representation of rural regions outside Kathmandhu, where the main training effort has been going on. 97% of Nepal is rural and 6 districts are contained in the Pokhara region and it includes many quite remote schools. It has relatively good road communications and some of the best contacts with schools and regional directorate of education were in that region. Access to sample schools was possible within a day, (if public transport could get fuel), so greatly reducing the cost of employing technical staff and limiting the overall time spread of test administration. (Given the limited time span of the study the longer it took to administer the baseline tests the less comparable would be the results of the study).

Selection of Teachers

The project leader in consultation with other training staff was asked during the first visit to select 16 teachers who had been trained in either the past or current Pokhara Inset courses.

The main selection criteria we required were:

i) The teachers were thought to be likely to implement their training, in its key characteristics.

ii) The teachers did not work in a school known to be exceptionally different from other sample schools.

The trainers were also asked to select 16 untrained teachers, with the aim of providing a roughly comparable control group. Initial selection was done according to best available local knowledge of the trainers and the local Regional Education Office.

Further selection criteria for both groups were:

i) Pupils and teachers in control (C) and experimental (E) groups should be as equivalent as possible in terms of language ability. Methods to ensure this as follows:

   a) Pupil equivalence:
   SLC results of (C) and (E) schools should be compared and schools with equivalent scores included in the study.

   b) Teacher equivalence:
   The Part 1 language test designed for the baseline study was administered to the teachers during the SEPELT trainers' initial visit in January. Teachers with widely disparate language levels were at this stage dropped from the study. This resulted in reducing the n in each group from 16 to 12.

This obviously meant that we were reducing the potential effect of improved English arising out of the training because our main concern in this study was to see if improved teaching methods made any difference to pupil language scores.

ii) There should be no special features in the school intakes which would bias the sampling, eg, extreme variations in parental income, school to school or rural versus urban.

iii) Access to the schools by technical staff should be both possible and welcome.

iv) The teachers should remain with their grade 8 class throughout grade 8 and should be likely to continue with the same pupils through grade 9.

v) There should be equivalent stability in pupil population in both (C) and (E) group schools; that is attrition rates should not differ markedly.

vi) All schools in the study should be well enough organised and run to ensure the efficient collection of test and observational data.

vii) There should be adequate facilities for testing, to minimise student copying

viii) As far as possible (C) group teachers should not receive informal "secondary training" during the period of the study, eg, by contact with trained teachers.

ix) It was essential that (C) teachers should not attend a training until late 1990.

x) Reaching each school should be possible in the period of the study, and very remote schools were to be excluded.

As far as possible an initial selection of 16 untrained and 16 trained teachers was made on the basis of these criteria with the expectation of some scaling down in sample size because much of the information necessary was simply not available in a documented form. The initial selection had to be made on the best available knowledge of the trainers and the Regional Education Office staff.

Because of the difficulties involved in selecting the sample we built an initial visit in January 1989 into the study in which the trainers were asked to collect data to determine the extent to which the above criteria were met. As a result the sample was cut down to twelve in the control and twelve in the experimental group. Two teachers subsequently left their posts and we thus finished up with a sample of 11 trained and 11 untrained teachers whose conditions are roughly comparable and on whom the study could be based.

A more careful screening of the schools as originally envisaged would obviously have been preferable. However, given the constraints in Nepal this was never feasible. In particular, the need to move the study forward at very short notice from its planned start date in March to January 1989 made these arrangements the best that could have been achieved.

Given the Nepali context and the nature of educational sampling in general we had no alternative than to base the study on an opportunity sampling. Random sampling was simply not feasible. We would have needed to select about three hundred teachers out of the total for upper secondary if this had been deemed necessary. This would have involved more time and expenditure than incurred in the rest of the project.

In summary, the study started with a selection of 16 control and 16 experimental schools which were considered likely to meet certain necessary conditions for inclusion. On the basis of screening visits in January 1989 8 of the schools which fell short of these criteria were removed from the study and we began the investigation with an n of 12 in each group. Since January, one trained and one untrained teacher left their schools and we eliminated their students' scores from the study.

Equivalence of the Groups

In the November 1989 visit we attempted to corroborate data on all the schools remaining in the sample in terms of: number of periods of language instruction received; continuation of teachers with the same group in year 9; additional training received by teachers in the untrained group; quality of pupil intake; overall academic performance as reflected by SLC results; student attrition.

Language Assessment of Teachers

During the November visit we were able to interview 18 of the 24 teachers and to assure ourselves that each had a base line language competence sufficient for them to teach the Nepali English curriculum in grades 8-10.

We managed to conduct tests on 18 of the 24 teachers in the study during the course of our visit. During the subsequent programme of visits the New Era staff administered all the tests with the exception of the oral to those not attending.

Method

Teachers (with the exception of the 6 not attending) were individually assessed on the basis of their performance in interviews, using the British Council's 9 band oral assessment checklist. To further determine their ability to teach English in the secondary system all teachers were given the students' tests which are based on grade 7 and 8 textbooks (the dictation and the gap filling). They were given an additional MCQ grammar test designed for University Entrance Language Proficiency screening in the UK.

In terms of their command of the structures and lexis in the books, there was little to choose between the two groups as was clearly shown in the dictation and gap filling tests. At this level they were both displaying a similar competence in the language. This is borne out by the t tests carried out on this data, where no significant difference can be shown between the control and experimental group teachers on the gap filling and dictation tests. Both

groups exhibited a high degree of competence in these textbook based tests and their level contrasts sharply with the rather poor estimates of teachers' language ability highlighted in Davies' 1984 report.

In the analysis done on the students' test scores (DICT1/2, RDG1/2, WRIT1/2), teachers 'scores on which we had complete data were taken into account when assessing the students' improved performance from February to November.

In the statistical analysis we looked at the contribution of the teacher language level to student test performance and found there to be a negligible effect. There is no indication that teachers' language ability had any noticeable effect on students' language scores.

In terms of the student samples we have to accept the non equivalence of the two groups on the basis of their initial test scores but note that the differences are not large. In any case the General Linear Models Procedure (GLIM analysis) we used to analyse the data took these differences into account.

In terms of size of class, school results and the number of hours spent there are small differences between the control and experimental groups but these are not statistically significant . Size of class (SIZE), number of hours tuition (HOURS), School SLC results both general (GENPASS) and in English (ENGPASS) made insubstantial contributions to test scores.

5 Contracting The New Era for Data Collection

A decision was taken in November 1988 that local technical staff should be contracted to collect test and process data on the grounds of economy and their Nepali field experience. The New Era research organisation was selected as the best source for such staff.

5.1 Observer training January 1989

A second visit to the Nepal project was made by Weir to conduct an observer training session for New Era staff who would be responsible for collecting data on the effects of training on pedagogical practice. A special training manual was produced by Roberts and Weir for this purpose. In addition it was necessary to familiarise these staff in the conduct of the language tests. This involved a briefing on the instructions for invigilation and the steps to be taken post testing.

The observer training would seem to have been effective from what emerged in the trialling in Katmandhu. After joint observations the schedules were compared and a reasonable degree of agreement was noted. Where any differences occurred these were the subject of later training sessions.

The best four out of the six New Era staff (ie those who had performed best in the training) were selected to carry out the subsequent observations and the testing.

5.2 Monitoring visit November 1989

A further monitoring visit was made by Weir and Roberts in November 1989 with the following objectives:

a) To visit schools jointly with New Era staff.
b) To review language test procedures with New Era staff.
c) To review collection of observational data, particularly checklists.

d) To monitor the selection of sample schools and teachers.
e) To make recommendations for future data collection based on b-d above.

The following were the outcomes of the visit

re a) : Sixteen schools were visited, and thirteen teachers were jointly observed by New Era staff and Wir/Roberts between 5.11.89 and 10.11.89.

re b) : Language test procedures were reviewed in an initial briefing meeting with New Era staff in Pokhara on 4.11.89.

: The administration of tests was subsequently monitored in 5 schools and found to be satisfactory.

re c) : Observation procedures and category interpretations were reviewed and agreed in the meeting and a joint observation held on 4.11.89.
: Subsequent joint observations were reviewed and discussed.

On the basis of post lesson comparisons, there appeared to be a satisfactory level of reliability between observations made by Weir and Roberts and New Era staff.

6 Summary of the Data Available

By the end of 1989 the following data were available for analysis.

Language assessments

: Students' language tests: 22 schools (11 trained, 11 untrained); after removing outliers we had 716 students' script.
: Teachers' language tests: 22 completed tests; 18 oral estimates

Interviews

: 18 interviews (9 untrained, 9 trained)

Process Descriptions

:	Observations	: 22 teachers, 69 observation forms
:	Teacher self report	: 20 teachers (10+10), 54 reports
:	Sample student work	: 20 teachers (10+10)

6.1 Test Data

The analysis was carried out using SAS and in particular the General Linear Models Procdure (GLIM). As a first step the outliers in the population were removed from the sample by plotting the scores on graphs for each of the three tests, dictation, reading and writing . Their status as outliers was determined by their extreme position on the plotted scattergram. Candidates in the first administration of the tests scoring more than 15 on the

dictation, or 24 on the gap filling, or 8 on the writing task, were removed from the sample as it was considered they were too dissimilar from the population we were interested in. This left us with an N of 343 students in the experimental group and 373 in the control group.

In all we had data on the performance of these two groups on the two sittings of the gap filling test (GAP 1 & GAP 2), the dictation (DICT 1 & DICT 2) and the writing (WRIT 1 and WRIT 2) In addition we were able to take into account the effect of a number of other variables on these test scores.

We had collected data on the teachers in the 2 groups on the same tests (DICT & READING) and on the grammar test (GRAM). The size of the classes attending the first test (SIZE) and the estimates of the number of hours of English each class had (HOURS) in the academic year 1989 were also available. The percentage of class time pupils spent talking in English (PUPENG) and the number of criterial features of training demonstrated by the teachers (FEATS) were also included in the analysis. Finally we have more limited data on the general pass rate of the schools in the SLC (GENPASS) and the English pass rate (ENGPASS) at the SLC.

6.2 Process Data

The observation schedule produces two quantifiable measures and supporting unquantified descriptions. The quantified data consists of :

a) Pupil English : a raw number of pupil English (PE) codings, against codings for all kinds of talk, which can then be expressed as a percentage[PENG].

b) Criterial features : checklist entries which identify trained teachers'typical activities (cats.2,3,4,6,7,8,9,10,11)[FEATS] and untrained teachers' typical activities (cats 1,5).

Both these measures can be aggregated for the comparison of control and experimental groups. In our study, GLIM analysis included the variables of pupil English [PENG] and criterial features [FEATS]. The unquantified data in the observation schedules [notes with checklist entries and whole lesson descriptions] were used to conduct internal validity checks, by identifying the consistency between descriptions, checklist entries, and codings.

The 54 self report lesson descriptions were analysed by categorizing reported activities, identifying those associated with untrained and trained teachers, and displaying their relative incidence in the two groups.

Samples of student work were not analysed, as an insufficient number matched either the lessons observed or teachers'self reports.

7 METHODOLOGICAL ISSUES ARISING FROM THE STUDY

7.1 LANGUAGE TESTS

The use of the same test at the beginning and end of treatment is open to the criticism that any improvement may be due to practice effect. Given that our purpose is to compare the performance of the two groups we might reasonably assume that the practice effect benefits both groups equally. There was an eight month gap between the two administrations and students did not know they would be taking the same tests again. If we take scores on the first test into account in the analysis of the second administration this

enables us to contrast gains made by the two groups. If there is a difference between the two groups in peformance on the second test administration it can be reasonally inferred that the training has had some effect. It would be imprudent, however, to try to isolate any specific training features as causes for observed changes in test scores. A cluster of associated variables result from training; and it is not possible to identify with any certainty the relative contribution of individual variables to outcomes.

More difficult to answer are questions relating to the worth of any differences that might emerge. This is particularly the case in a non skills based test when one has to convert a quantitative score gain on discrete linguistic items into an interpretation. The judgement to be made on size of gain is in itself problematic when dealing with quantitative scores. If the gain is large the interpretation is better grounded.

In the case where the gain is relatively small one might wish to consider this from a longer term perspective. One might point to possible exponential as against linear gain in future test scores, given the possibility of initial inertia and old habits. If the teaching of English is to take place over a number of years any differences between groups might be magnified in future years. This of course argues that in studies of this type monitoring over a period of years would be required.

There is a critical need to ensure that some of the tests used in these studies are fair in content to both control and experimental group in order to make valid and fair comparisons. In addition there is a need to try to develop tests which are sensitive to particular features of training to confirm differential treatments. By definition the latter tests are not fair to both groups.

In situations such as Nepal where both groups are using the same course book and working towards the same final school examination, developing tests which were fair to both groups was possible by basing the test items on materials and activities common to both groups.

The greater problem lay in devising tests to reflect differences in pedagogical practice. We had to identify differences in treatment and develop tests which would measure effects on student performance. As we have indicated above this was problematic for two reasons. First, there is a general difficulty in establishing with any certainty causal relationships between pedagogical treatment and learning outcomes. Secondly, as we discuss below, there is a difficulty in identifying what the criterial features will be at the stage of the implementation of training rather than in the training itself.

There is a further problem where a study is designed to measure gain over a period of time in setting items at a suitable level of difficulty. If the items are too easy then a ceiling effect would quickly ensue and prevent any long term comparison. If the items are too difficult then it may be that they would be insensitive to gain even over an extended period. There needs to be a balance of items in terms of difficulty. We attempted to do this by basing the tests on Book 7 which all students had completed at the start of year 8 and also book 8 which they would have completed by the end of the first year of the baseline study. Given a very low start rate, poor previous learning experiences, and a limited number of hours of English, however, even items based on units covered may be beyond the reach of most of the students. In addition they may already be severely demotivated and this could interfere with the effects of any enhanced treatment.

It seemed sensible (in the absence of clear implicational items) to include tests with a large number and range of items in the first administration. In this way any differences would have a better chance of emerging and would enhance the reliability of the results obtained. There is always a danger that long tests might discourage students but observations did not suggest such an effect in this situation. It may be that statistical modelling through item response theory might be of some help in long term studies in the future.

It may well be that tests such as gap filling and dictation, because they focus on specific linguistic items may be testing constructs which take a long time to develop in learners. There is some suggestion in second language acquisition research that gains in linguistic competence may take a longer time to appear in comparison with skills development and performance. It may be the case that had we been able to develop practical tests of say spoken language ability, gains in test scores might have been clearly marked. This

is an area which is in urgent need of research. The practical problems in testing skills such as spoken interaction cannot be ignored, however, and the limitations this imposes on evaluation studies are evident.

A final practical constraint is the length of time that the data even from a small scale study such as this takes to collect and process. At a conservative estimate it involved around 150 person days. This would be quite a sizeable chunk of a project member's time that would need to be allocated to evaluation.

7.2 PROCESS DATA

There are a number of reasons for treating the process data in this particular study with some caution and there are potential lessons for future evaluations to be learned.

1. The number of observations was not really sufficient to be sure of giving an adequate picture of teachers' customary practice. The conventional view is that about six visits are needed for this. Had New Era staff been able to obtain two observations per visit, the data would have been greatly improved. The requirement to complete tests within the shortest possible time, and in November, the imminence of the end of the school year, militated against this.

   This raises the question of the extent to which insiders should be involved, albeit in a formative evaluation role. For summative decisions the case for involving outsiders stands and it might be necessary to have a small number of outsider observations than none at all.

2. The reliability of observations is recognised as being greatly improved by the use of paired observers. For cost and logistical reasons this was not an option.

3. For reasons of extreme difficulty in access and travel, joint observations by New Era staff and the external evaluators were very much on an opportunity basis. In such circumstances one might not be able to conduct joint observations as one would want.

4. Quantification of process data is dependant on the identification of adequate units. In this study coding was based upon the recognition of utterance units rather than, for example, arbitrary time units. As a result coding boundaries were dependant on interpretation. For example when teachers or pupils repeat themselves, make false starts, or give one word responses, or during continuous speech such as in teacher explanation, it is possible that different observers might identify different numbers of utterances. In the case of question and answer exchanges, where counting speech utterance units is considerably easier, greater agreement can be expected. Ideally, it would be worthwhile to train observers in the the use of time unit coding, but considerable resources would have to be available.

5 The performance of observed teachers isoften influenced by "impression management". It may be that teachers provided "lessons to order" in which features of training would appear. The high occurrence of Pupil English in E group lessons and the high occurrence of Teachers' English rather than Nepali in C group lessons may over-represent the norm. Triangulation of data is the necessary strategy to validate observations. In this study a number of sources were used including teacher self report (see Appendix 2) and feedback forms. Our self report data suggested that criterial differences continued to be exhibited in unobserved lessons. The returns to an insider post course evaluation

questionnaire in 1988 presented a similar picture. Trained teachers reported some use of oral drills but in other areas, such as writing activities, they admit to extremely restricted application of training.

In spite of these limitations in the observational data, if the differences in criterial indicators between C and E groups are marked then they are likely to reflect real differences in classroom experiences for students.

7.3 LONG RANGE EVALUATION

Our experiences in Nepal highlighted the problems of outsider evaluation through long distance monitoring. As a rule evaluators working in this way are only able to spend a limited amount of time in the field. By necessity they have to work through others both in terms of setting up the study and in implementing it. Making practical arrangements such as meetings with teachers, visits to schools, organising transport and petrol, selecting a sample for the study, are all that much more difficult at a distance especially when internal communication in the country concerned is problematic.

In the absence of an extended field based feasibility study, outsiders also have to rely on insiders to provide them with information on which to construct the evaluation instruments. In our case we relied on the trainers to provide us with our information on the criterial features of training which they considered would be implemented by the experimental group. A feasibility study would have enabled us to scrutinise classroom practices as would attendance at a wider range of the training sessions. It would have enabled us for example to omit any concern with writing in Nepali classrooms from the study as subsequent experience demonstrated that because of the low importance of this in the school leaving examinations and because of time constraints on teachers, this activity was absent from most classroom practice. In terms of cost and time required, however, an extensive initial survey by outsiders may not be funded.

It does seem that this is a strong argument in favour of increased systematic internal monitoring by project staff. This would promote a more accurate definition of the categories of information for use in outsider evaluations. However, at crucial points outsider monitoring would be necessary as there is a risk of contamination of the data collected by personnel with an investment in the success of the project.

In our experience, contracting technical staff to conduct observations had unexpected benefits. We now realise that this means of collecting observational data is likely to produce an explicit analysis of criterial variables and the use of low inference criteria in observation along with less structured methods. As insider observers do not have the same need for expliciness or objectivity in producing observational data there is a real risk that their findings might not be meaningful to a wider audience.

7.4 ASPECTS OF SAMPLING

The results of our small scale survey are at best suggestive rather than conclusive as random sampling was not an option. In the event we were able to sample 11 out of the 1080 trained teachers.

We attempted as far as we were able to control for a variety of variables which might contaminate the results: class size, school leaving results, language level of teachers, attrition rates and number of hours of instruction received. By using the General Linear Models procedure to carry out the statistical analysis we were able to take account of these variables when determining the effect of treatment on language test scores.

Ethical problems arise in this type of non equivalent control group design. In particular there must be some concern about the anomalous position of the control group teachers. A small retainer was paid to encourage their participation in the study (preparing self report forms, attending meetings etc.). The effect of this has been to defer their training. Also we were acutely aware that in entering their classes our role was very much that of outsiders looking for evidence of deficit. Careful consideration needs to be given to their interests.

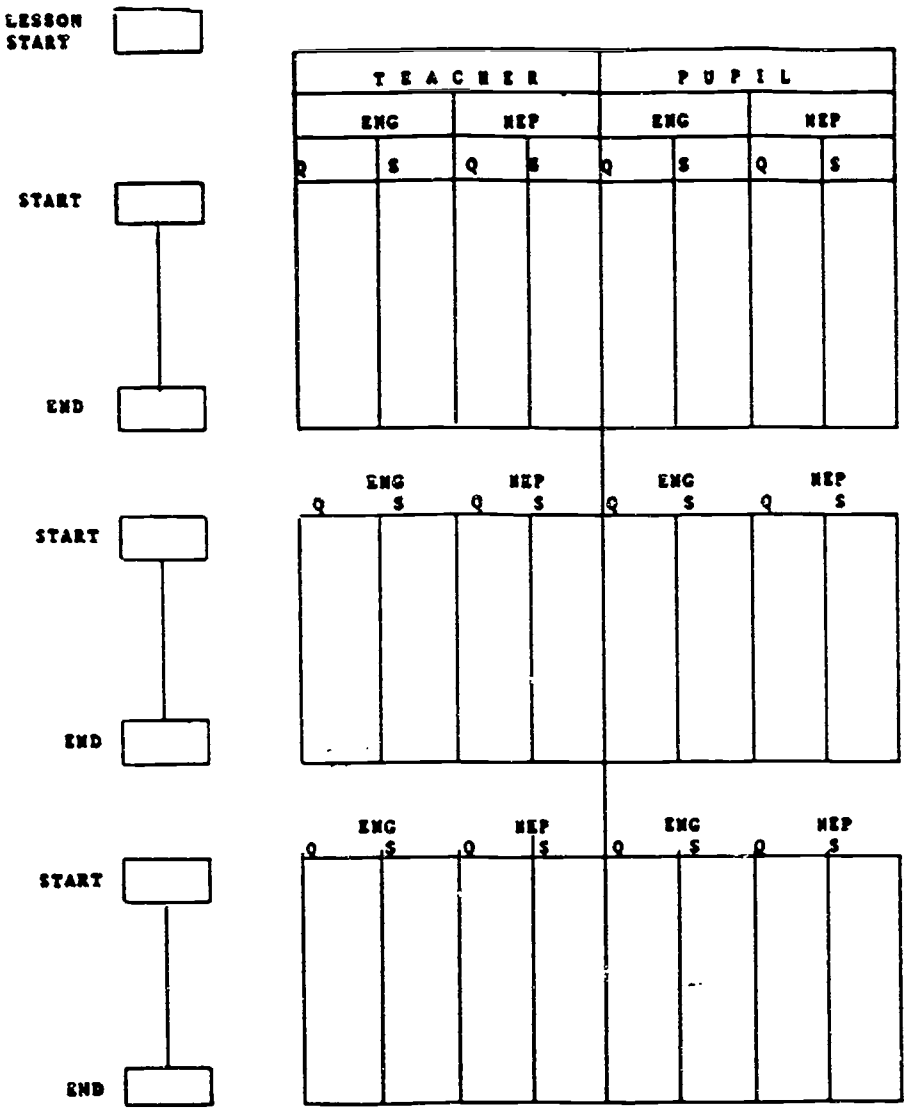
APPENDIX ONE

SEPELT BASELINE STUDY OBSERVATION FORM

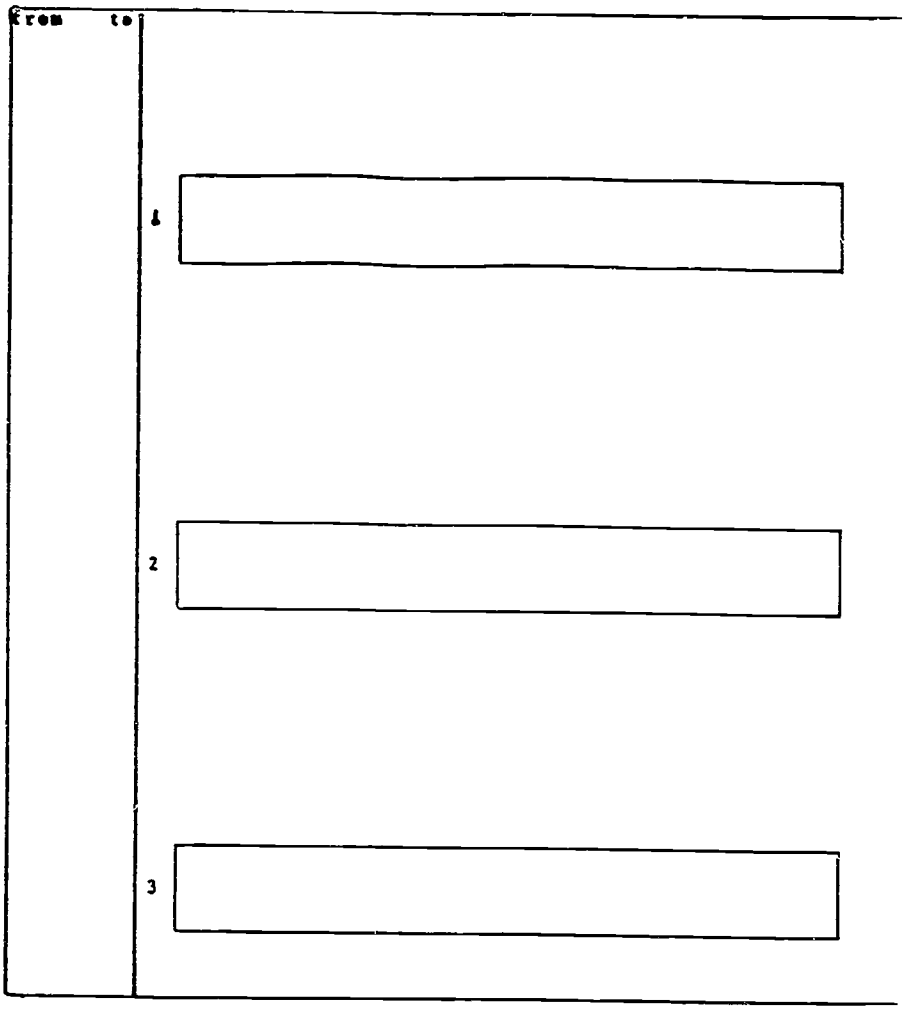
OBSERVER	
TEACHER	
SCHOOL	
CLASS	Number of pupils present :
LESSON/ PAGE	
DATE	

Teacher's Lesson Outline

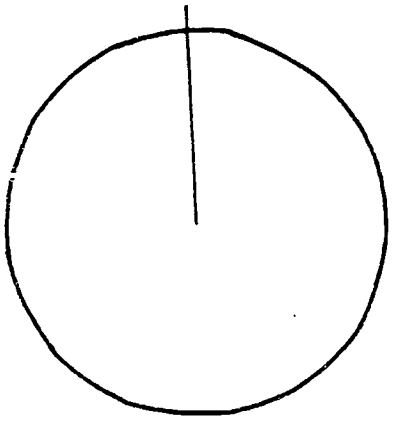
APPENDIX A (Cont'd)



APPENDIX A (Cont'd) STAGES OF THE LESSON



	Teacher explained grammar mainly in Nepali		Pupils produced original sentences in English
	Teacher practised language in situations mainly in English		Pairwork or groupwork used
	Teacher gave models in English		Teacher gave listening practice
	Oral drills used,in English		Extra practice
	Pupils did written exercises by copying		Notes on extra practice:
	Pupils did guided writing exercise		
	Comprehension questions asked in English		



19

APPENDIX TWO

YOUR NAME Buddhi Prasad Sharma

LESSON DESCRIPTION

Please fill in the form, describing what happened in a recent, ordinary lesson: what you did, what the pupils did, the exercises you used, the work the children produced and so on. If you can, please attach an example of one pupil's work. Thank you.

Date May 15, 1989

Class VIII A
Sts. 69

WHAT I DID IN THE LESSON	WHAT THE PUPILS DID IN THE LESSON
- Textbook item: paragraph writing on 'How I Make Halwa' page 73	
- Eight instructional sentences from page 73 exercise C were written on the card board beforehand.	
- After reading the sentences, pictures and real objects were shown for difficult words.	
- Each sentence modelled twice e.g. Put a cupful of ghee in a pan. I put a cupful of ghee in a pan	Repeated by the pupils in chorus then individually.
- pupils were asked to write the same model sentences omitting the numbers so as to make a paragraph	- Class work was done by the pupils
- while writing the paragraph, necessary check up of the sentences was made	- Best paragraphs were read by the pupils out in the class

How.
Some instructional sentences were given to write a paragraph on 'How I Make a Cup of Tea'